Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije

Veliki jezikovni modeli za digitalno humanistiko / Large language models for digital humanities: LLM4DH

Izroček projekta/ Project Deliverable

Načrt ravnanja z raziskovalnimi podatki / Data management plan

Št. Izročka / Deliverable No.	7.2.1
Datum oddaje / Delivery date	7. 3. 2025 (Verzija 1.2)
Vrsta / Nature[1]:	R
Nivo diseminacije / Disemination level[1]:	PU
Delovni Sklop / Work package	WP7 Management and Dissemination T7.2 Ethics, Quality, Risk, and Management
Avtor, Vodilni partner / Author, Lead partner	Marko Robnik-Šikonja, UL FRI
Sodelujoči partnerji / Contributing partners	Špela Arhal Holdt, Marko Bajec, Ciril Bohak, Slavko Žitnik, UL FRI Senja Pollak, Simon Krek, Nikola Ljubešić, Matej Martinc, Marko Pranjić, IJS Darja Fišer, Andrej Pančur, Kaja Dobrovoljc, INZ Polona Tratnik, IRRIS Aleš Završnik, Saša Krajnc, IK Katja Dobrovoljc, Špela Vintar, UL FF Simon Dobrišek, UL FE Darinka Verdonik, Andrej Žgank, UM FERI

[1]Nature: R = Report, P = Prototype, D = Demonstrator, O = Other (specify)
[1]Dissemination level PU = Public
CO = Confidential, only for members of the consortium (including ARIS)

Projekt financira ARIS – Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije, št projekta GC-0002, izbranega v okviru Javnega razpisa za sofinanciranje Gravitacije.

Informacijska tabela projekta / Project information card	
Akronim projekta / Project acronym	LLM4DH
Naslov projekta/ Project title	Veliki jezikovni modeli za digitalno humanistiko / Large language models for digital humanities
Spletna stran / Website	https://www.cjvt.si/llm4dh/
Št projekta / Project No.	GC-0002
Št Pogodbe / Contract no.	Pogodba o (so)financiranju znanstvenoraziskovalne dejavnosti v letu 2024 št. 1000-24-0510 (UNI FRI)
Razpis / Call	Javni razpis za (so)financiranje Gravitacije, št: 5100-70/2024-6 z dne 7.5.2024 in Sprememb javnega razpisa za (so)finanfiranje Gravitacije, št. 5100-70/2024-11 z dne 29.5.2024.
Koodrinator projekta / Project coordinator	Marko Robnik-Šikonja
Trajanje projekta / Project duration	1.10.2025 – 30.9.2027, 36 mesecev / 36 months
Povzetek projekta / Abstract of the project	Jezikovne tehnologije igrajo vse pomembnejšo vlogo na številnih področjih znanosti. Z bliskovitim napredkom velikih jezikovnih modelov se ti pozicionirajo kot osrednja tehnologija umetne inteligence z daljnosežnim vplivom ne le na znanstveno raziskovanje ampak tudi na širši družbeni razvoj in celotno družbo. Veliki jezikovni modeli, kljub svojim velikim zmožnostim, izkazujejo vrsto pomanjkljivosti, kot so nekonsistentni in napačni odgovori, velike računske zahteve, šibko poznavanje jezikov z manj viri, neprilagojenost za nekatere pomembne naloge in domene ter slabosti v razumevanju družbe, etike in človekovih potreb. Projekt bo tehnologijo velikih jezikovnih modelov izboljšal z različnimi znanji in podatki, jim zmanjšal računsko zahtevnost pri govornih tehnologijah in jim omogočil dostop do zanesljivih zunanjih informacij. Izboljšave v temeljnih zmožnostih velikih jezikovnih modelov bodo vodile do prebojnih dosežkov na več področjih digitalne humanistike: v jezikoslovju, leksikografiji, izobraževanju, analizi družbenih pojavov in sodobne zgodovine, političnih vedah, folkloristiki in pravu. Language technologies are increasingly important in many areas of science. With the rapid advances in large-scale language models, they are positioning themselves as a core AI technology with far-reaching impact on scientific research, wider societal development, and society. Despite their great capabilities, large language models exhibit several shortcomings, such as inconsistent and incorrect answers, high computational demands, weak knowledge of less-resourced languages, poor performance in some important tasks and domains, and weaknesses in understanding society, ethics, and human needs. The project will improve the technology of large language models by enriching them with diverse knowledge and data, using them to improve the performance of speech technologies, and giving them access to reliable external information. Improvements in the core capabilities of large language models will lead to breakthroughs in several areas of the digital

	humanities: linguistics, lexicography, education, analysis of social phenomena and contemporary history, political sciences, folkloristics, and law.
Vodilni partner projekta/ Lead project partner	Univerza v Ljubljani, Fakulteta za računalništvo in informatiko (**UL FRI)**
Sodelujoči partnerji / Participating project partners	Institut "Jožef Stefan" (**IJS**) Inštitut za novejšo zgodovino (**INZ**) Inštitut IRRIS za raziskave, razvoj in strategije družbe, kulture in okolja (**IRRIS**) Inštitut za kriminologijo pri Pravni Fakulteti v Ljubljani (**IK**) Univerza v Ljubljani, Filozofska fakulteta (**UL FF**) Univerza v Ljubljani, Fakulteta za elektrotehniko (**UL FE**) Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko (**UM FERI**)
Opis aktivnosti 7.2.1 / Description of the Action 7.2.1	Pripravili bomo Načrt ravnanja z raziskovalnimi podatki (NRRP), ki bo vseboval informacije o: obravnavi raziskovalnih podatkov med projektom in po njegovem zaključku, kateri podatki bodo zbrani, obdelani in/ali ustvarjeni, katere metodologije in standardi bodo uporabljeni, ali bodo podatki deljeni/dostopni javnosti ter kako bodo podatki upravljani in ohranjeni (tudi po zaključku projekta). Načrt bo posodobljen z razvojem projekta. Upravitelj podatkov bo spremljal ustvarjene nabore podatkov in zagotovil, da bodo v skladu s postopki, določenimi v NRRP. **Načrt je pripravljen po navodilih v obrazcu NRRP ARIS.** A Data Management Plan (DMP) will be developed that consists of information on the handling of research data during and after the end of the project, what data will be collected, processed and/or generated, which methodology and standards will be applied, whether data will be shared/made open access and how data will be curated and preserved (including after the end of the project). This will be updated throughout the project development. The data manager will keep track of generated data sets and ensure they will fit the procedures in the DMP. **The plan follows the instructions provided in the ARIS DMP form.**

Zgodovina dokumenta / Document history			
Datum / Date	Verzija	Avtor / Author	Komentar / Comment
3.2.2025	1.0	Marko Robnik-Šikonja Maja Zupančič Justin	Gre za živ dokument, ki ga bomo nadgrajevali tekom projekta. This is a live document that will be updated continually during the project.
4. 2.2025	1.1	Marko Robnik-Šikonja Maja Zupančič Justin	Dokument smo opremili z glavo in spremno informacijo. Dokument smo poslali partnerjem, da so dodali potrebne informacije. We have added a header and accompanying information to the document.
6. 3. 2025	1.2	Marko Robnik-Šikonja Maja Zupančič Justin	Končni pregled in konsolidacija dokumenta. Umik navodil.

KAZALO VSEBINE

1. Načrt ravnanja z raziskovalnimi podatki po navodilih ARIS

NAČRT ZA RAVNANJE Z RAZISKOVALNIMI PODATKI

OBRAZEC ARIS

Large Language models for digital humanities / Veliki jezikovni modeli za digitalno humanistiko
LLM4HD

(Razpis: Gravitacija GC-0002, ARIS)

Deliverable 7.2: Data management plan
Verzija: 1.2
Vodja raziskovalnega projekta: prof. dr. Marko Robnik-Šikonja
Datum priprave: 7. 3. 2025

Odgovorni člani projektne skupine za pripravo in vzdrževanje dokumenta:
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko (**UL FRI)**: Marko Robnik-Šikonja / Špela Arhal Holdt / Marko Bajec / Ciril Bohak / Slavko Žitnik
Institut "Jožef Stefan" (**IJS**) : Senja Pollak / Simon Krek / Nikola Ljubešić / Matej Martinc / Marko Pranjić
Inštitut za novejšo zgodovino (**INZ**) /: Darja Fišer / Andrej Pančur / Kaja Dobrovoljc
Inštitut IRRIS za raziskave, razvoj in strategije družbe, kulture in okolja (**IRRIS**): Polona Tratnik
Inštitut za kriminologijo pri Pravni fakulteti v Ljubljani (**IK**): Aleš Završnik / Primož Križnar
Univerza v Ljubljani, Filozofska fakulteta (**UL FF**): Katja Dobrovoljc / Špela Vintar
Univerza v Ljubljani, Fakulteta za elektrotehniko (**UL FE**): Simon Dobrišek
Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko (**UM FERI**): Darinka Verdonik / Andrej Žgank

Obrazec je namenjen pripravi načrta za ravnanje z raziskovalnimi podatki (NRRP) za raziskovalne projekte, ki jih (so)financira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS), kot je določeno v 4. členu Uredbe o izvajanju znanstvenoraziskovalnega dela v skladu z načeli odprte znanosti (Uradni list RS, št. 59/23).

Raziskovalni podatki so opredeljeni kot zapisi o dejstvih (številčni podatki, besedilni, zvočni in slikovni zapisi), ki predstavljajo osnovno podlago za znanstveno raziskovanje in ki v okviru znanstvene skupnosti veljajo kot ustrezno sredstvo za preverjanje veljavnosti raziskovalnih spoznanj.

Prosimo vas, da izpolnite spodnji obrazec NRRP in ga posredujete ARIS **najkasneje v šestih mesecih od začetka izvajanja raziskovalnega projekta**. Priporočljivo je, da NRRP med izvajanjem raziskovalnega

projekta po potrebi redno pregledujete in posodabljate. V primeru sprememb posodobljen NRRP priložite vmesnemu in zaključnem poročilu o rezultatih raziskovalnega projekta.

Pregled vsebine NRRP:

0. Splošne informacije
1. Povzetek in opis raziskovalnih podatkov
2. Shranjevanje in varnostno kopiranje podatkov
3. Zagotovitev podatkov na način FAIR
   - 3.1 Zagotavljanje najdljivosti podatkov (F)
   - 3.2 Zagotavljanje dostopnosti podatkov (A)
   - 3.3 Zagotavljanje interoperabilnosti podatkov (I)
   - 3.4 Zagotavljanje ponovne uporabe podatkov (R)
4. Etični in pravni vidiki
5. Drugi raziskovalni rezultati
6. Finančna sredstva

Uporabljene kratice:

- ADP – Arhiv družboslovnih podatkov
- GDPR – Splošna uredba o varstvu podatkov
- IT – informacijska tehnologija
- RO – raziskovalna organizacija
- ZVDAGA – Zakon o varstvu dokumentarnega in arhivskega gradiva ter arhivih
- ZVOP-2 – Zakon o varstvu osebnih podatkov

Navodilo za izpolnjevanje

Obrazec izpolnite tako, da vsebino vnašate v celice v skrajnem desnem stolpcu oz. tam označite eno od ponujenih možnosti. V teh celicah so sedaj v sivi barvi pisave *navedeni razlage oz. navodila za vnos ustreznih podatkov in opisov. To pomožno besedilo lahko po vnosu vsebine izbrišete.*

0	Splošne informacije	
0.1	**Šifra projekta**	GC-0002 (Javni razpis za sofinanciranje Gravitacije)
0.2	**Naziv projekta**	Veliki jezikovni modeli za digitalno humanistiko / Large language models for digital humanities LLM4DH
0.3	**Šifra vodje projekta**	15295
0.4	**Ime in priimek vodje projekta**	Marko Robnik Šikonja
0.5	**Ime in priimek osebe, ki je v RO zadolžena za podporo pri ravnanju z raziskovalnimi podatki**	V okviru projekta LLM4DH koordinira ravnanje z raziskovalnimi podatki prof. dr. Marko Robnik Šikonja. Na projektu LLM4DH je udeleženih osem RO. V nadaljevanju so navedene osebe oziroma komisije, ki so zadolžene za podporo pri ravnanju z raziskovalnimi podatki v okviru posamezne RO. **UL FRI:** Vodja UL FRI »Komisije za rokovanje z raziskovalnimi podatki (KRRP)« je izr. prof. dr. Veljko Pejović. **IJS:** Znanstveno-informacijski center (ZIC) IJS, skrbnica podatkov je Helena Klančnik **INZ**: Filip Dobranić **IRRIS:** prof. dr. Polona Tratnik **IK**: prof. dr. Aleš Završnik **UL FF:** asist. dr. Matej Petrič **UL FE:** znan. svet. dr. Urša Opara Krašovec **UM FERI**: dr. Jernej Kusterle
0.6	**Interna pravila RO za ravnanje z raziskovalnimi podatki**	**UL FRI:** UL FRI ima interni dokument z naslovom »Komisija za rokovanje z raziskovalnimi podatki (KRRP)«, ki opredeljuje postopek oddaje vloge z raziskavo, ki vključuje zbiranje in uporabo potencialno občutljivih podatkov. Ravno tako ima UL FRI izdelane usmeritve za pripravo načrta ravnanja z raziskovalnimi podatki (NRRP) in obrazec za rokovanje z z raziskovalnimi podatki za doktorske študente, ki sledi načelom FAIR in načelu Horizon EU glede odprtega dostopa. **IJS: /** **INZ**: / **IRRIS:** / **IK**: / **UL FF: /** **UL FE: /** **UM FERI**: interna pravila so v postopku priprave
0.7	**Verzija NRRP**	1.2
1	Povzetek in opis raziskovalnih podatkov	
1.1	**Ali boste pri projektu ponovno uporabili že**	**x Da** ☐ Ne **Obrazložitev**:

	obstoječe podatke predhodnih raziskav?	Obrazložitev podajamo ločeno po raziskovalnih sklopih projekta: **WP1 Improving LLMs** **T1.1 Improving LLMs with linguistic data / T1.2 Improving LLMs with cross-lingual data / T1.3 Improving multimodal models** **Namen**: za razvoj in izboljšave velikih jezikovnih modelov **Obstoječi podatki:** • Digitalna slovarska baza CJVT UL (SBZ) • Besedilni korpusi in baze Gigafida, Metafida, Wiktionary, BabelNet, z DBPedio povezani WordNeti • Slovenski prevod llava_v1_5_mix665k podatkovne množice **WP2 LLMs for Linguistics and Knowledge Management** **T2.1 Neural spell- and grammar checking / T2.2 LLMS for effective lexicography / T2.3 Advanced grammatical analysis of multilingual corpora** **Namen:** Kvalitativne in kvantitativne analize avtentičnih jezikovnih napak v slovenskih besedilih. Vir primerov za strojno spreminjanje zgledov iz jezikovno ustreznih v jezikovno neustrezne. **T2.1 Neural spell- and grammar checking** **Obstoječi podatki** • Podatki besedilnih korpusov z označenimi jezikovnimi napakami: Lektor (Popič 2014), KOST (Stritar Kučuk 2023), Šolar (Arhar Holdt & Kosem 2024) in Šolar-Eval (Gantar et al. 2024). • Referenčni korpusi, zlasti Gigafida (Krek et al. 2020). • Korpusna zbirka Universal Dependenies 2.15 (Zeman et al. 2024) **WP3 LLMs for Spoken Language** **T3.1 Efficient spoken language data collection and annotation / T3.2 Semantic and pragmatic speech processing / T3.3 Quality assessment of speech data / T3.4 LLMs for domain-specific speech recognition** **Namen:** Uporaba podatkov za učenje modelov ASR ter modelov za semantično in pragmatično označevanje besedil; označevanje že obstoječih podatkov z na novo naučenimi modeli **Obstoječi podatki**: • ROG (http://hdl.handle.net/11356/1992) • GOS 2.1 (http://hdl.handle.net/11356/1863) • Artur (http://hdl.handle.net/11356/1776) **WP4 Advanced Technologies for DH** **T4.1 Named entity graphs / T4.2LLMs for diachronic analysis / T4.3 Multimodal models for image analysis / T4.4 LLMs and RAG for contradiction detection** **Namen:** za razvoj pristopov ekstrakcije imenskih entitet in povezav, izgradnjo grafov znanja, pristopov za odkrivanje semantičnih sprememb **Obstoječi podatki** • Slovenian and multilingual historical corpora • multilingual historical corpora

		• Semantic change detection dataset for Slovenian 1.0 • sPeriodika 1.0, a corpus of Slovenian historical periodicals (1771–1914) • ParlaMint • CLASSLA-web.sl • Carniolan Provincial Assembly corpus Kranjska 1.0 • Parliamentary corpus of first Yugoslavia (1919-1939) yu1Parl 1.0 • Ilustrirani slovenec – vse izdaje (predvsem od 1924 naprej) • Ilustrirani glasnik – vse izdaje • Revija Vesna – vse izdaje • Ciciban **WP5 DH Challenges** **T5.1 Historical newspapers and ideological analyses / T5.2 LLMS for folkloristics / T5.3 LLMs and RAG for legal domain** **Namen:** Za raziskovanje odnosov med ljudmi, kraji in organizacijami, predvsem pan-slovanskih, jugoslovanskih in slovenskih identitet, odnose teh pojmov do zgodovinskih centrov moči in vprašanje kako se te identitete mešajo s prej omenjenimi**.** **Obstoječi podatki** • sPeriodika 1.0, a corpus of Slovenian historical periodicals (1771–1914) (T5.2), • rokopisi, traktati in etnografska besedila (spomini, dnevniki, revije, potopisi itd.) različnega evropskega izvora (španski, francoski, nemški, angleški, italijanski, slovenski in hrvaški) od 15. do 19. Stoletja (T5.2), • Digitalizirana revija Ciciban (pdf) (T5.2), • Digitaliziran arhiv slovenskih ljudskih pravljic ZRC SAZU (T5.2), • digitalizirani stari tiski ali ilustracije (približno 500) o ritualu reševanja sporov (T5.2), • CLASSLA-web.sl • Carniolan Provincial Assembly corpus Kranjska 1.0 • Parliamentary corpus of first Yugoslavia (1919-1939) yu1Parl 1.0 **WP6 Evaluation and Understanding of LLMs** **T6.1 Figurative language and pragmatics benchmarks / T6.2 Spoken language understanding benchmark / T6.3 Bias detection in LLMs and ASR / T6.4 Knowledge-based explanation of LLMs** **Namen:** za razvoj in izboljšave velikih jezikovnih modelov **Obstoječi podatki**: • Metaphor and Sarcasm Scenario Test (Adachi et al. 2004) • WOW and the WAX dataset of association explanations (Liu et al. 2022)

		• Training corpus of spoken Slovenian ROG 1.0 (http://hdl.handle.net/11356/1992) • EEC bias evaluation dataset (Kiritchenko & Mohammad 2018)
1.2	**Katere vrste podatkov boste ustvarili oz. ponovno uporabili in v katerih formatih bodo shranjeni?**	**WP1 Improving LLMs** **T1.1 Improving LLMs with linguistic data / T1.2 Improving LLMs with cross-lingual data / T1.3 Improving multimodal models** • Surova besedila iz SBZ, format txt • Grafi znanja iz SBZ, format json • Slikovno-besedilni nabor (dataset), ki vsebuje slike s slovenskimi napisi, format json Gre za standardne formate, ki so v rabi pri razvoju LLM. **WP2 LLMs for Linguistics and Knowledge Management** **T2.1 Neural spell- and grammar checking** • Sintetično pripravljeni primeri z jezikovnimi napakami in težavami: gre za besedilne podatke v tabelarični obliki, uporabili bomo format CSV, ki je tipična izbira za takšne podatke na CLARIN.SI • Novi dataseti za evalvacijo strojnega pregledovanja: gre za besedilne podatke z jezikoslovnimi oznakami, uporabili bomo format CoNLL-U, ki je tipična izbira za takšne podatke na CLARIN.SI **/ T2.2 LLMS for effective lexicography / T2.3 Advanced grammatical analysis of multilingual corpora** • Množica z navodili za prilagajanje modela, format txt • Množica za evalacijo uspešnosti modelov, format txt **WP3 LLMs for Spoken Language** **T3.1 Efficient spoken language data collection and annotation / T3.2 Semantic and pragmatic speech processing / T3.3 Quality assessment of speech data / T3.4 LLMs for domain specific speech recognition** T3.1 in T3.2 Razširjen učni govorni korpus ROG-Art, dodatnih 5 ur govora, skupno 10 ur – XML + WAV T3.1 in T3.2 Razširjen SloBench ASR, dodatna 1 ura govora, skupno 4 ure – XML + WAV T3.1 Zbirka posnetkov zasebnega govora Govorjena slovenščina - WAV T3.3 Avdio baza posnetkov govora iz javno dostopnih virov – WAV **WP4 Advanced Technologies for DH** **T4.1 Named entity graphs / T4.2LLMs for diachronic analysis / T4.3 Multimodal models for image analysis / T4.4 LLMs and RAG for contradiction detection** T4.2 • sPeriodika 1.0, a corpus of Slovenian historical periodicals (1771–1914) - extracting poverty subcorpus - besedilni • ParlaMint - extracting poverty subcorpus - besedilni

		• Semantic change detection dataset for Slovenian 1.0 - besedilni in številčni T4.4: • Surova besedila iz virov (a) PISRS.si, (b) Uradnega lista, (c) Sodbe ustavnega sodišča in (d) SodnaPraksa.si. Besedila bodo shranjena v izvornem formatu spletnih vmesnikov, t.j. TXT in JSON. • T4.4: Ustvarili bomo povezave med posameznimi deli pridobljenih dokumentov, kot učno množico za učenje kontradikcij. T4.3 • Za Ilustrirani Slovenec, Ilustrirani glasnik, Ciciban in Revijo vesno bomo uporabili prvotne pdf-je, potencialno pretvorjene v png format, obogatene z avtomatsko pridobljenimi meta informacijami, kot so koordinate slik v pdf-jih in besedni opisi slik v formatu json **WP5 DH Challenges** **T5.1 Historical newspapers and ideological analyses / T5.2 LLMS for folkloristics / T5.3 LLMs and RAG for legal domain** • T5.2: tekstovni podatki sPeriodika 1.0, rokopisi, traktati in etnografska besedila, digitaliziran arhiv slovenskih ljudskih pravljic ZRC SAZU, revija Ciciban – surovi tekst v TXT formatu • T5.3: Uporabili bomo podatke iz T4.4 in izdelali podatkovno množico za odgovarjanje na vprašanja za izbrane primere uporabe. **WP6 Evaluation and Understanding of LLMs** **T6.1 Figurative language and pragmatics benchmarks / T6.2 Spoken language understanding benchmark / T6.3 Bias detection in LLMs and ASR / T6.4 Knowledge-based explanation of LLMs** • T6.1: prevodi in prilagodbe Metaphor and Sarcasm Scenario Test in WOW and the WAX dataset of associations, jsonl • T6.2: razširjen SloBench ASR, dodatna 1 ura govora, iz T3.1 in T3.2, wav, xml, jsonl • T6.3: prilagojeni EEC bias evaluation dataset (Kiritchenko & Mohammad 2018), jsonl, razširjen SloBench ASR, jsonl • T6.4: ontologije in grafi znanja, tsv, jsonl
1.3	**Kakšen je namen ustvarjanja, zbiranja oz. ponovne uporabe podatkov in njihova povezava s cilji projekta?**	**WP1 Improving LLMs** **T1.1 Improving LLMs with linguistic data / T1.2 Improving LLMs with cross-lingual data / T1.3 Improving multimodal models** Novo ustvarjeni podatki bodo vmesna faza pri učenju in izboljšanju delovanja velikih jezikovnih modelov in velikih multimodalnih modelov. **WP2 LLMs for Linguistics and Knowledge Management** **T2.1 Neural spell- and grammar checking / T2.2 LLMS for effective lexicography / T2.3 Advanced grammatical analysis of multilingual corpora**

		Novi podatki, ustvarjeni na T2.1 bodo uporabljeni za učenje in evalvacijo strojnega slovničnega pregledovanja pisnih besedil v sodobni slovenščini. Novo ustvarjeni podatki bodo v sklopu T2.3 uporabljeni za razvoj in evalvacijo metod za sintezo slovnično označenih besedilnih korpusov. **WP3 LLMs for Spoken Language** **T3.1 Efficient spoken language data collection and annotation / T3.2 Semantic and pragmatic speech processing / T3.3 Quality assessment of speech data / T3.4 LLMs for domain-specific speech recognition** T3.1 in T3.2 Novo ustvarjeni podatki bodo v projektu uporabljeni za učenje in evalvacijo modelov razumevanja človeškega govora T3.1 in T3.3 Novo ustvarjeni podatki bodo v projektu uporabljeni za razvoj in evalvacijo metod za avtomatsko klasifikacijo posnetkov **WP4 Advanced Technologies for DH** **T4.1 Named entity graphs / T4.2LLMs for diachronic analysis / T4.3 Multimodal models for image analysis / T4.4 LLMs and RAG for contradiction detection** T4.1: Izdelava orodij za ekstrakcijo imenskih entitet in relacij ter izgradnjo grafov znanja. T4.2: Izdelava podpornega orodja za odkrivanje pomenskih sprememb T4.4: Izdelava podpornega orodja za iskanje kontradikcij v pravnih dokumentih, aktih, sodbah, zakonih in njihovih predlogih. **WP5 DH Challenges** **T5.1 Historical newspapers and ideological analyses / T5.2 LLMS for folkloristics / T5.3 LLMs and RAG for legal domain** T5.1: Izdelava vizualizacij za grafe znanja iz T4.1. T5.2: izdelava digitalnega korpusa iz obstoječih podatkovnih baz, anotiranje zbranih podatkov in metapodatkov ter analiziranje podatkov z analitičnimi metodami T5.3: Izdelava specializiranega pravnega pomočnika za izbrane primere uporabe, ki bodo definirani tekom projekta oz. po vrednotenju rezultatov v T4.4. **WP6 Evaluation and Understanding of LLMs** **T6.1 Figurative language and pragmatics benchmarks / T6.2 Spoken language understanding benchmark / T6.3 Bias detection in LLMs and ASR / T6.4 Knowledge-based explanation of LLMs** T6.1: izdelava testnih množic za figurativni jezik in pragmatično uporabo, namen je evalvacija tekstnih velikih jezikovnih modelov T6.2 izdelava testnih množic za razumevanje govora, namen je evalvacija govornih velikih jezikovnih modelov T6.3 izdelava testnih množic za odkrivanje pristranskosti tekstnih, govornih in slikovnih velikih jezikovnih modelov T6.4 izdelava testnih množic za razlago velikih jezikovnih modelov na podlagi znanja

1.4	**Kakšna je pričakovana velikost podatkov, ki jih nameravate ustvariti oz. ponovno uporabiti?**	☐ 0–10 GB ☐ 10–100 GB ☐ 100–1000 GB **X >1000 GB** Veliki jezikovni modeli, veliki multimodalni modeli za svoje učenje in prilagajanje potrebujejo velike količine podatkov. Celoten projekt bo ustvaril veliko količino takih podatkov v različnih formatih in za različne namene. Celotno količino ustvarjenih podatkov ocenjujemo na nekaj 10 TB.
2	**Shranjevanje in varnostno kopiranje podatkov**	
2.1	**Kje bodo podatki med izvajanjem projekta shranjeni in varnostno kopirani?**	**UL FRI** uporablja lasten varen strežnik za podatke izven javne domene. Za javno dostopne podatke raziskovalci UL FRI uporabljajo institucionalni podatkovni strežnik, Arnesov podatkovni strežnik, Sling podatkovni strežnik ter hrambo Onedrive in Google Drive. Vsi ti podatkovni strežniki in repozitoriji so varnostno kopirani. Razvita izvorna koda bo deponirana v odprte repozitorije, kot so GitHub, GitLab in Sourceforge. Odprti LLM-ji so deponirani v repozitorij HuggingFace. Novi odprti nabori podatkov bodo deponirani v repozitorij CLARIN.SI. Pri **UMB** bodo novo ustvarjeni podatki shranjeni na lokalnih diskih UMB. Ob koncu projekta bodo deponirani tudi v repozitorij CLARIN.SI. **IJS** za podatke uporablja lastni strežnik NextCloud. Razvita izvorna koda bo deponirana v odprte repozitorije, kot so GitHub/GitLab, modeli v repozitorij HuggingFace, novi odprti nabori podatkov pa v repozitorij CLARIN.SI. Drugi projektni partnerji uporabljajo infrastrukturo UL FRI, UMB, and IJS.
2.2	**Kako boste izbrali podatke za dolgoročno hrambo?**	Cilj projekta je spodbujanje odprte znanosti. Kadar koli je to mogoče, bodo pridobljeni ali ustvarjeni podatki trajno javno dostopni in deponirani v infrastrukturo Clarin.si. Za avtorsko zaščitene podatke, npr. korpus Ciciban, bomo upoštevali pogodbene obveznosti med institucijo prejemnico in imetnikom avtorskih pravic, vendar ti podatki ne bodo objavljeni. Sledili bomo 57.a (besedilno in podatkovno rudarjenje), 57.b (besedilno in podatkovno rudarjenje za namene znanstvenega raziskovanja) in 57.c (znanstveno raziskovanje) členom Zakona o avtorskih in sorodnih pravicah (Url. RS št 16/07 – uradno prečiščeno besedilo, 68/08, 110/13, 56/15, 63/16 – ZKUASP, 59/19 in 130/22), ki opredeljujejo pravico do hranjenja podatkov in njihovo rabo za znanstvene namene.

2.3	**Ali bodo podatki shranjeni v zaupanja vrednem repozitoriju?**	**X Da** ☐ Ne Obrazložitev: Vsi novo nastali podatki, ki jih je mogoče javno objaviti, bodo deponirani v repozitoriju CLARIN.SI. Nastali veliki jezikovni modeli bodo deponirani v odprtem repozitoriju HuggingFace.
3.	**Zagotovitev podatkov na način FAIR**	
3.1	**Zagotavljanje najdljivosti podatkov (F)**	
3.1.1	**Ali bodo podatki označeni s trajnim identifikatorjem (PID)?**	**X Da** ☐ Ne Obrazložitev: Vsi vnosi v repozitorij Clarin.si dobijo trajni identifikator (handle). Znanstvene objave v ciljnih revijah in na odprtih repozitorijih, kot je ArXiv, dobijo identifikator DOI.
3.1.2	**Kateri metapodatki bodo ustvarjeni in kateri metapodatkovni standardi bodo pri tem upoštevani?**	Repozitorija Clarin.si in HuggingFace, ki ju bomo uporabljali za trajno shranjevanje ustvarjenih podatkov in modelov, definirata svoj nabor metapodatkov, ki jim je potrebno slediti in ki jih bomo ustvarili. Repozitorij clarin.si izmenjuje metapodatke z metarepozitorijem VLO čez protokole CMDI in OLAC. Repozitorij je registriran v OpenAIRE in re3data.
3.1.3	**Ali bodo metapodatki vsebovali ključne besede za izboljšanje najdljivosti in možnosti ponovne uporabe?**	**X Da** ☐ Ne Obrazložitev: Uporabljali bomo ključne besede in kategorije, ki jih predpisujejo repozitoriji, kamor bomo odlagali ustvarjene podatke.
3.2	**Zagotavljanje dostopnosti podatkov (A)**	
3.2.1	**Ali bodo vsi podatki odprto dostopni?**	☐ Da **x Ne** Obrazložitev: Odprtodostopni bodo vsi podatki, za katere tega ne prepovedujejo imetniki avtorskih pravic in kjer je to tako določeno v pogodbah z njimi (npr. korpus Ciciban). Omejitve bodo tudi pri dostopnosti govornih podatkov, kjer bo to potrebno zaradi varovanja osebnih podatkov govorcev (govor kot biometrični osebni podatek). Izjema pri odprti dostopnosti je tudi del podatkov pri evalvacijskih množicah (T2.1 in T2.3), ki mora ostati skrit, da evalvacija lahko poteka metodološko ustrezno in v skladu z mednarodno raziskovalno prakso.
3.2.2	**Kdaj bodo podatki odprto dostopni in za koliko časa?**	Raziskovalni podatki bodo načeloma dostopni ob objavah v repozitorijih in v znanstvenih publikacijah.

		Izjema pri odprti dostopnosti je tudi del podatkov pri evalvacijskih množicah (T2.1 in T2.3), ki mora ostati skrit, da evalvacija lahko poteka metodološko ustrezno in v skladu z mednarodno raziskovalno prakso. WP3: Izdelani govorni viri bodo odprto dostopni ob koncu projekta in časovno ne bodo omejeni. Dostopnost podatkov ne bo časovno omejena.
3.2.3	**Na kakšen način bo v primeru omejitev pri uporabi omogočen dostop do podatkov med izvajanjem projekta in po njegovem zaključku?**	**WP1 Improving LLMs** Korpusa Gigafida in Metafida: Potreben bo podpis izjave o varovanju avtorskih pravic z UL. T1.3. Večina podatkov za treniranje slikovno-jezikovnih modelov ter vse podatkovne množice za testiranje modelov bodo prosto dostopne. V primerih, ko določene podatkovne množice za treniranje modelov ne bodo prosto dostopne zaradi prepovedi imetnikov avtorskih pravic, bomo hranili podatke na internih strežnikih. **WP2 LLMs for Linguistics and Knowledge Management** **T2.1 Neural spell- and grammar checking / T2.2 LLMS for effective lexicography / T2.3 Advanced grammatical analysis of multilingual corpora** Za T2.1 in T2.3 ne predvidevamo omejitev dostopa do podatkov. Izjema je del podatkov pri evalvacijskih množicah, ki mora ostati skrit, da evalvacija lahko poteka metodološko ustrezno. **WP3 LLMs for Spoken Language** **T3.1 Efficient spoken language data collection and annotation / T3.2 Semantic and pragmatic speech processing / T3.3 Quality assessment of speech data / T3.4 LLMs for domain specific speech recognition** Omejitve pričakujemo pri govornih virih zaradi upravljanja osebnih podatkov. Predvidoma bodo posnetki govora, zlasti zasebnega, na voljo pod pogojem, da bo uporabnik posnetkov sporočil svoj kontakt (e-naslov) in opredelil lastno zakonito podlago za obdelavo posnetkov kot samostojen upravljavec podatkov. Na podlagi GDPR bo lahko tudi po zaključku projekta posamezen posnetek izločen iz govornih virov, če bi tako na podlagi GDPR zahteval govorec na posnetku. **WP4 Advanced Technologies for DH** **T4.1 Named entity graphs / T4.2LLMs for diachronic analysis / T4.3 Multimodal models for image analysis / T4.4 LLMs and RAG for contradiction detection** T4.2. Pri večini podatkov za diahrone analize ne bo omejitev, metodologija bo javno dostopna. V primerih, ko podatkovne množice za posamične diahrone analize niso prosto dostopne, bomo hranili podatke na internih strežnikih in

		bodo podatki dostopni v okviru dogovorov z zainteresiranimi raziskovalci. T4.3. Pri večini podatkov za multimodalno analizo slik ne bo omejitev, metodologija bo javno dostopna. V primerih, ko podatkovne množice za posamične analize niso prosto dostopne, bomo hranili podatke na internih strežnikih in bodo podatki dostopni v okviru dogovorov z zainteresiranimi raziskovalci. T4.4: Pri pravnih besedilih bodo uporabljeni javni viri in ne bo potrebe po varovanju. Besedila, ki bodo avtorsko zaščitena (npr. pravni učbeniki), bodo vsebovani v osnovnem jezikovnem modelu, ki ne bo del tega projekta. **WP5 DH Challenges** **T5.1 Historical newspapers and ideological analyses / T5.2 LLMS for folkloristics / T5.3 LLMs and RAG for legal domain** T5.3: Enako kot pri T4.4. **WP6 Evaluation and Understanding of LLMs** **T6.1 Figurative language and pragmatics benchmarks / T6.2 Spoken language understanding benchmark / T6.3 Bias detection in LLMs and ASR / T6.4 Knowledge-based explanation of LLMs** Velik del ustvarjenih podatkov bodo evalvacijske množice, pri katerih morajo pravilni odgovori ostati skriti, da evalvacija lahko poteka metodološko ustrezno in v skladu z mednarodnimi raziskovalnimi praksami.
3.2.4	**Ali bo za dostop do podatkov oz. njihovo branje potrebna dodatna dokumentacija oz. informacija o ustrezni programski opremi?**	☐ Da **x Ne** Obrazložitev: Vsi podatki so shranjeni v odprtih formatih, za njihovo rabo ni potrebna specializirana programska oprema.
3.3	**Zagotavljanje interoperabilnosti podatkov (I)**	
3.3.1	**Katere geslovnike oz. šifrante boste uporabili pri pripravi podatkov in metapodatkov?**	Uporabljali bomo standardno kategorizacijo repozitorijev CLARIN.SI in Hugging Face. Repozitorij CLARIN.SI poganja sistem LINDAT, ki temelji na tehnologiji CLARIN DSpace in ustrezni metapodatkovni shemi. Uporabili bomo standarde za jezikoslovno označevanje slovenščine, ki so popisani na https://wiki.cjvt.si/shelves/jezikoslovno-oznacevanje-korpusov.
3.3.2	**Ali boste primorani uporabiti manj poznane ali lastne geslovnike oz. šifrante?**	☐ Da **x Ne** Obrazložitev: Povsod bomo uporabljali znane in odprte standarde zapisovanja.
3.4	**Zagotavljanje ponovne uporabe podatkov (R)**	

3.4.1	**Na kakšen način boste zagotovili dokumentacijo, potrebno za ponovno uporabo podatkov?**	Sledili bomo navodilom repozitorijev Clarin.si in HuggingFace, ki zahtevajo opis metapodatkov in podrobnejši opis podatkov. Kjer je mogoče, bomo ustvarjenim virom dodali povezave na znanstvene publikacije.
3.4.2	**Ali bodo vaši podatki javno dostopni in licencirani v skladu z odprto licenco CC0, da bo s tem omogočena čim širša ponovna uporaba?**	☐ Da **X Ne** Obrazložitev: Večina podatkov bo dostopna pod odprtimi licencami CC-BY in CC-BY-ATTR 4.0. Veliki jezikovni modeli bodo odprtodostopni pod licenco Apache-2, programska oprema pa pod licenco Apache-2, MIT ali GNU-3. Bolj zaprte licence bodo uporabljene v primeru, kjer bo to potrebno zaradi upravljanja osebnih podatkov ali avtorskih pravic.
3.4.3	**Kakšne postopke zagotavljanja kakovosti podatkov boste uporabili?**	Izdelane podatkovne množice bodo podvržene najprej temeljitemu pregledu avtorjev in vodje delovnega sklopa, nato pa še internemu pregledu kakovosti na nivoju uporabnikov ustvarjenih podatkov.
4.	**Etični in pravni vidiki**	
4.1	**Ali obstajajo etična ali pravna vprašanja, ki bi lahko vplivala na deljenje podatkov?**	**X Da** ☐ Ne Obrazložitev: Pravna vprašanja, ki zadevajo uporabo podatkov za velike jezikovne modele, naslavljamo v sodelovanju s projektom NOO PoVeJMo s pomočjo zunanje pravne svetovalke, dr. Maje Jančič Bogataj. Za govorne podatke je ugotovitev pravnih strokovnjakov, da ni splošne ovire za ureditev skupne dostopnosti posnetkov na portalu Clarin. Upravljanje osebnih podatkov v govornih virih naslavljamo s pomočjo najetih zunanjih pravnih svetovalcev pri Aphaia.
4.2	**Ali boste med izvajanjem projekta obdelovali oz. hranili osebne podatke?**	☐ Da **X Ne** Obrazložitev: Projekt obdeluje pretežno odprte podatke. Za ostale vrste podatkov sledi Splošni uredbi o varstvu podatkov (GDPR). Pridobljeni viri, ki bi lahko vsebovali osebne podatke, bodo predhodno anonimizirani. Pri zajemanju govornih podatkov bomo obdelovali, hranili in distribuirali govor kot biometrični osebni podatek. V ta namen smo izdelali izjavo o varstvu osebnih podatkov in izjavo za privolitev v snemanje in obdelavo osebnih podatkov kot tudi strinjanje s pogoji prispevanja posnetkov. Za pravno svetovanje so najeti zunanji pravni strokovnjak pri Aphaia.
4.3	**Ali bodo med projektom ustvarjene oz. ponovno**	☐ Da **X Ne**

	uporabljene posebne vrste osebnih podatkov?	Obrazložitev: Projekt obdeluje pretežno odprte podatke. Za ostale vrste podatkov sledi Splošni uredbi o varstvu podatkov (GDPR). Pridobljeni viri, ki bi lahko vsebovali osebne podatke, bodo predhodno anonimizirani. Pri zajemanju govornih podatkov bomo obdelovali, hranili in distribuirali govor kot biometrični osebni podatek. Pri tem bomo podatke upravljali s pomočjo zunanjih pravnih strokovnjakov pri Aphaia.
4.4	**Kako boste uredili lastništvo avtorskih pravic in pravic intelektualne lastnine podatkov, ki jih boste ustvarili ali ponovno uporabili?**	Lastništvo avrtorskih pravic in pravic intelektualne lastnine konzorcijski partnerji projekta LLM4DH urejajo s konzorcijsko pogodbo. Konzorcijski partnerji (pogodbene stranke) se zavezujejo k medsebojnemu varovanju zaupnih informacij in pravic intelektualne lastnine ter avtorskih pravic, pri čemer zagotavljajo, da bodo: • skrbno in preudarno ravnali z vsemi informacijami, pridobljenimi od pogodbenih strank pri izvajanju operacije, ki preprečuje njihove razkritje, objavo ali razširjanje brez njihovega predhodnega soglasja; • spoštovali pravice intelektualne lastnine ali licenc iz naslova kakršnokoli blagovne znamke, avtorske pravice ali patenta, ki ga pogodbena stranka že ima ali ga bo pridobil oziroma ga bo nadzorovala kasneje. Intelektualna lastnina oz. znanje, ki je nastalo pred datumom pričetka izvajanja projekta in ki je razvito izven projekta, pripada pogodbeni stranki, ki je/so sodeloval/i pri nastanku te intelektualne lastnine oz. znanja (obstoječe znanje). Pogodbene stranke ohranijo vse znanje in pravice intelektualne lastnine, ki je nastala in bila zaščitena oziroma so bili postopki zaščite sproženi pred pričetkom izvajanja projekta ali ki je razvita izven projekta. Pogodbena stranka ima pravice dostopa do intelektualne lastnine in znanja druge pogodbene stranke, če te rezultate potrebuje za izvedbo svojega dela v okviru projekta. Takšen dostop se dovoli brezplačno. Pogodbena stranka ima pravice dostopa do intelektualne lastnine in znanja druge pogodbene stranke, če te rezultate potrebuje za izkoriščanje lastnih rezultatov, pri čemer se takšen dostop dovoli pod poštenimi in razumnimi pogoji, kar se uredi s posebno pogodbo. Zahteva za dostop po tem odstavku se lahko poda največ eno leto po zaključku projekta, če ni dogovorjeno drugače. Rezultati so last pogodbene stranke, ki jih je dosegla oziroma ustvarila. Rezultati pomenijo vse materialne in nematerialne rezultate dejavnosti, kot so podatki, znanje ali informacije,

		ustvarjeni v okviru projekta, ne glede na njihovo obliko ali vrsto in ne glede na to, ali se lahko zaščitijo ali ne, ter vse njim pripadajoče pravice, vključno s pravicami intelektualne lastnine. Če pogodbene stranke rezultate dosežejo skupaj in če njihovega posameznega prispevka k skupnemu rezultatu ni mogoče določiti ali razdeliti, so ti rezultati v skupni lasti tistih pogodbenih strank, ki so jih ustvarile. V tem primeru skupni lastniki sklenejo poseben dogovor, s katerim podrobneje uredijo razdelitev pravic intelektualne lastnine in pogoje uporabe skupnih rezultatov.
5.	**Drugi raziskovalni rezultati**	
5.1	**Ali boste poleg podatkov ustvarili ali ponovno uporabili tudi druge raziskovalne rezultate?**	**x Da** ☐ Ne Obrazložitev: Ustvarjenih bo več digitalnih orodij, kot so veliki jezikovni modeli in programska koda. Odprt dostop in uporaba bosta omogočena po načelih FAIR. Zagotovili bomo, da bodo digitalni predmeti in drugi raziskovalni podatki objavljeni, najdljivi, dostopni, interoperabilni in uporabni.
6.	**Finančna sredstva**	
6.1	**Kakšni bodo stroški ravnanja s podatki in drugimi rezultati projekta po načelih FAIR in kako bodo kriti?**	Načeloma so stroški varne hrambe podatkov in zagotavljanje ustrezne podatkovne infrastrukture znatni. Raziskovalne organizacije jih pretežno pokrivajo iz javnih sredstev za raziskovalno dejavnost, iz projektne režije in sponzorskih sredstev. Javne raziskovalne infrastrukture, kot je Clarin.si, so financirane iz državnega proračuna. Projekt sam teh stroškov načeloma ne nosi, razen posredno skozi režijske stroške.
6.2	**Kdo bo odgovorna oseba za ravnanje z raziskovalnimi podatki pri projektu?**	prof. dr. Marko Robnik-Šikonja

Uporabljeni viri:

- *Anotirana predloga načrta za ravnanje s raziskovalnimi podatki za projekte Obzorja Evropa.* CTK UL. Dostopno na: https://dirrosdata.ctk.uni-lj.si/raziskovalni-podatki/nacrt-ravnanja-z-raziskovalnimi-podatki/.
- Bezjak, Sonja (ur.) (2024). *Spoznaj FAIR: Priročnik o odprti znanosti v Sloveniji*. Univerza na Primorskem. Dostopno na: https://www.hippocampus.si/ISBN/978-961-293-328-9.pdf.
- *Horizon Europe Data management plan template*. Dostopno na: https://www.openaire.eu/images/Guides/HORIZON_EUROPE_Data-Management-Plan-Template.pdf.
- *NWO Template Data management plan*. Dostopno na: https://www.nwo.nl/en/research-data-management.

Podpisani izjavljam, da so z vsebino Načrta za ravnanje z raziskovalnimi podatki seznanjeni vsi konzorcijski partnerji ter se z njo strinjajo.
Podpis:

_____
(Zastopnik oz. pooblaščena oseba raziskovalne organizacije, ki je pri projektu v vlogi vodje konzorcija)

ARIS obrazec:
Verzija dokumenta: 1.0
Številka: 6311-94/2024-1
Datum: 15. 10. 2024

2. Reference

ARHAR HOLDT, Špela, KOSEM, Iztok. Šolar, the developmental corpus of Slovene. *Language resources and evaluation*. 2024, vol. , no. , str. 1-27, tabele. ISSN 1574 - 20X. https://link.springer.com/article/10.1007/s10579-024-09758-4, DOI: 10.1007/s10579-024-09758-4. [COBISS.SI-ID 204228867]

GANTAR, Polona, BON, Mija, GAPSA, Magdalena, ARHAR HOLDT, Špela. Šolar-Eval : evalvacijska množica za strojno popravljanje jezikovnih napak v slovenskih besedilih. *Jezik in slovstvo*. [Tiskana izd.]. 2023, letn. 68, št. 4, str. 89-108, ilustr. ISSN 0021-6933. https://journals.uni-lj.si/jezikinslovstvo/article/view/16577, Digitalna knjižnica Slovenije - dLib.si, DOI: 10.4312/jis.68.4.89-108. [COBISS.SI-ID 187559683],

POPIČ, Damjan, 2014: Revising Translation Revision in Slovenia. Tamara Mikolič Južnič, Kaisa Koskinen, Nike Kocijančič Pokorn (ur.): New Horizons in Translation Research and Education 2. Joensuu: University of Eastern Finland. 72–89. http://epublications.uef.fi/pub/urn_isbn_978-952-611657-0/urn_isbn_978-952-61-1657-0.pdf.

STRITAR KUČUK, Mojca, 2023: Prvi korpus slovenščine kot tujega jezika KOST 1.0. Špela Arhar Holdt, Simon Krek (ur.): Razvoj slovenščine v digitalnem okolju. Ljubljana: Založba Univerze v Ljubljani. 93–117.

KREK, Simon, ARHAR HOLDT, Špela, ERJAVEC, Tomaž, ČIBEJ, Jaka, REPAR, Andraž, GANTAR, Polona, LJUBEŠIĆ, Nikola, KOSEM, Iztok, DOBROVOLJC, Kaja. Gigafida 2.0: the reference corpus of written standard Slovene. V: CALZOLARI, Nicoletta (ur.). *LREC 2020: Twelfth International Conference on Language Resources and Evaluation: May 11-16, 2020, Marseille, France.* Paris: ELRA - European Language Resources Association. 2020, str. 3340-3345. http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf

Zeman, Daniel; et al., 2024, *Universal Dependencies 2.15*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-5787.